面向企业微博的客户细分框架*

陈东沂[1,3] 周子程[1] 蒋盛益[1] 王连喜[2] 吴佳林[1]
[1](广东外语外贸大学信息学院 广州 510006)
[2](广东外语外贸大学图书馆 广州 510420)
[3](顺丰科技有限公司 深圳 518000)

摘要: 【目的】为有效解决微博客户特性的表示问题，以更好地实施企业微博客户细分。【方法】借助微博平台上客户的个人和社会关系特性，利用客户及其好友的自定义标签表示客户的特性，采用基于非负矩阵分解的文本聚类方法，提出一种面向企业微博的客户细分框架。【结果】实验结果表明，基于非负矩阵分解的方法取得约 86.130%的 asw 指标平均值，远远超出基于 K-means 和层次聚类的方法。【局限】只通过融合微博客户个人及其关注好友的标签表示微博客户特性的方法不能够全面刻画客户特征。【结论】能够为企业微博客户细分中的客户特性的表示、细分、评价及结果可视化等问题提供参考和借鉴。

关键词: 客户细分 微博营销 文本聚类 非负矩阵分解
分类号: TP391 G35

1 引 言

微博(Microblog)等社会化媒体的快速发展深刻改变了企业与客户、客户与客户之间的沟通和互动方式。微博具有信息传播快、互动性强、实时分享等特点，充分利用这些特点进行社会化营销能够帮助企业改善品牌形象，提高知名度，从而扩大市场份额，这使得微博营销成为企业社会化营销的重要手段，而客户细分是微博营销的重要基础。

自美国学者温德尔·史密斯于 20 世纪 50 年代中期提出客户细分的概念以来，客户细分研究引起了政府机构、工业界和学术界的广泛关注。目前，客户细分研究在企业客户关系管理中发挥着重要作用。然而，传统的企业客户细分方法存在局限性，新兴的营销方式和电子化平台使得传统的客户细分方法面临挑战。在社会化营销领域，传统方法难以有效表示客户的特性，面对海量社会化媒体数据，分析效果差强人意。

在已有研究的基础上，本文以企业微博为研究对象，以文本聚类技术为手段，研究面向企业微博的客户细分框架，探索客户细分在社会化营销中的应用。微博用户的社会关系及其兴趣爱好等标签对客户特性的表示具有重要意义。前期研究中，许多学者针对微博平台客户的社会关系特性，融合客户及其微博好友的自定义标签，从客户个人和社会特性两方面生成客户特性描述文本，并利用文本分类技术识别出微博平台的潜在客户；实验结果表明潜在客户识别准确率可以达到 86%左右[1]。在此基础上，本文利用文本聚类技术，并结合内部评价方式和标签云可视化方法，提出一种面向企业微博的客户细分框架；通过对不同行业的企业官方微博的粉丝数据进行分析，对比不同文本聚类方法的效果，结果表明 K-means 和层次聚类等传统算法倾向于粗略划分企业微博客户以致聚类效果不佳，而本文框架所采用的基于非负矩阵分解的文本聚类能有效处理高维文本数据和语义聚类，该框架有利

通讯作者：周子程，ORCID：0000-0001-9164-9494，E-mail：ziceweek@gmail.com。

*本文系国家自然科学基金项目"面向微博公共事件的反向社会情绪识别及演化分析研究"(项目编号:61572145)、广东省科技计划项目"广东省企业竞争情报信息提取及态势推理机制研究——以汽车行业为例"(项目编号:2015A030401093)和广东大学生科技创新培育专项资金项目"微博用户生成内容挖掘及其在微博营销领域的应用研究"(项目编号:308-GK151019)的研究成果之一。

于发现更有意义的客户群。

2 相关研究

2.1 客户细分相关研究

社会化媒体的快速发展为企业与客户提供了全新的互动交流平台，基于社会化媒体的客户关系构建和营销策略成为企业长远发展的制胜点[2]。客户细分成为社会化媒体营销最重要、也是企业管理者最为关注的方面之一[3]。面向微博的客户细分可以帮助企业快速分析客户群特性，开拓营销渠道，从而降低企业的营销成本、增加利润。

传统的企业客户细分方法主要有聚类和分类两种[4-6]。由于分类方法需要大量有标注的训练数据，要求企业对已有的客户资料及客户群特征有较好的认识，因此在实际应用中分类并不是主流的客户细分方法。聚类分析不需要标注的训练数据，只需对数据进行相似度计算以自动划分，是目前使用较多的细分方法，能有效发现企业客户群特征。Rajagopal 使用聚类技术识别零售业中的高收益、高价值和低风险的客户[4]；Lefait 等根据客户购买行为信息，提出一种基于聚类的客户细分框架以帮助企业细分客户群[5]；Wu 等提出不同的客户矩阵模型，融合聚类技术，发现客户的不同特性[6]。然而这些研究大多是传统行业里的客户细分应用，在分析方法和分析对象等方面存在一定的局限性，难以延伸应用到社会化营销领域。从方法的角度看，聚类或分类都需要在特定的条件下进行。基于划分的聚类方法大多需要指定划分的数目，而分类需要大规模的标注训练数据，也涉及参数设置的问题。传统方法主要分析客户的人口统计信息、消费特征等数值属性。但这些属性往往不能有效地表示客户特性，难以从兴趣爱好等方面刻画客户特性，导致细分的效果较差；同时，由于难以确保从社会化媒体平台获取的统计特征的真实性，直接运用传统方法效果并不好。

国内外针对社会化媒体平台上的企业客户细分研究成果尚不多见。国外学者探索了 Twitter 上的用户分类研究，包括政治立场分类、地域划分、性别预测和角色分类等[7-10]，但并没提及客户细分。国内微博发展起步较晚，更缺乏相关的研究。据笔者调查，目前国内外并没有直接的面向企业微博的客户细分研究。

2.2 文本聚类算法和非负矩阵分解简述

文本聚类是依据“同类文档相似度较大，不同类文档相似度较小”的假设，通过计算不同文档间的相似度将不同文档归到不同类别的过程。已有方法通过向量空间模型 VSM 和词频–逆文档频率 TF-IDF 权重计算，解决文本向量化表示的问题；进而，采用基于划分或层次等聚类算法，计算文本间的相似度实现聚类。常见的聚类算法有基于划分的、层次的、密度和网格的聚类算法等，比较经典且广泛运用的是 K-means 和层次聚类方法[11-13]。尽管文本聚类可以通过传统的聚类算法来实现，但这个过程仍存在问题：文本特征通常呈现高维和稀疏性，影响了文本聚类的效果；传统的聚类方法较少考虑文本语义，聚类结果难以直观呈现最终的聚类效果。

非负矩阵分解法(Non-negative Matrix Factorization, NMF)因 Lee 等于 1999 年发表在 *Nature* 的一系列研究成果而引起了学术界的关注[14]。非负矩阵分解可简述为：对于任意的非负矩阵 A，寻找非负矩阵 W 和非负矩阵 H，使得满足 A=W*H，从而将非负矩阵 A 分解为左右两个非负矩阵 W、H 的乘积。因此矩阵 A 中的某一列向量可以解释为对左矩阵 W 中所有列向量(称为基向量)的加权和，而权重系数为右矩阵 H 中对应列向量中的元素。这种表示具有直观的语义解释，反映了人类“局部构成整体”的思维。NMF 通过寻找降维表示和非负元素的矩阵分解形式的特点使其在实际的领域中得到了广泛的应用，比如文本挖掘、图像处理和生物信息处理等。借鉴 NMF 在文本分析的成功应用[15-19]，本文将 NMF 引入到对企业微博客户标签文本的处理中，发挥其处理高维数据的优势，并通过捕获文本语义信息实现文本聚类。

3 面向企业微博的客户细分

微博上蕴含了丰富的社会关系信息。微博用户间的关注关系说明用户之间存在真实的社会关系、或存在相似的兴趣爱好等特性、或对所关注用户分享的信息感兴趣。通常当用户关注某个企业微博时，说明该用户可能已经是该企业的客户，想继续了解企业的产品或服务；也可能是该企业的潜在客户，对企业的产品或服务感兴趣，但还没产生购买行为；还有小部分可能不是现有客户或潜在客户，而是企业的员工或竞

争对手，他们也可能是企业的客户，由于他们是业内人士，其个人和社会特性在一定程度上反映了企业的特征和客户的共性。因此，本文假定企业微博账号的粉丝为该企业的现有客户或潜在客户，有着相似的产品或服务需求，可以从不同侧面来描述他们的生活、职业和兴趣爱好等特征，这也使得聚类技术能够较好地发现这种潜在的模式。换言之，面向企业微博的客户细分问题可以看作对企业官方微博的粉丝进行细分的过程，可形式化为无监督的聚类问题。

3.1　面向企业微博的客户细分框架

本文结合企业官方微博的粉丝及其微博好友自定义的标签文本信息，将文本聚类技术应用到微博平台的客户细分研究中，提出一种面向企业微博的客户细分框架，如图 1 所示：

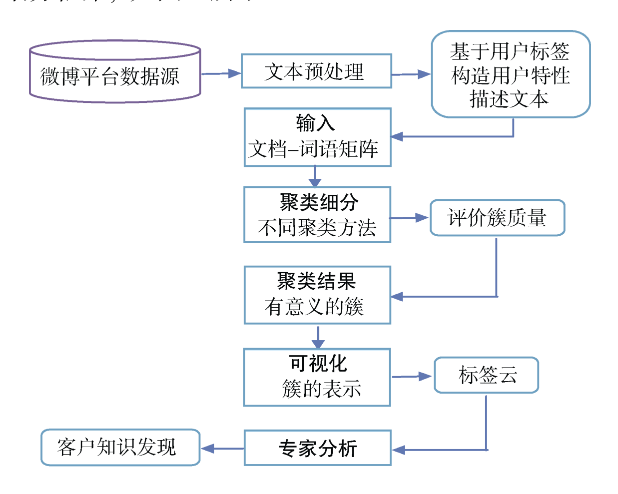


图 1　面向企业微博的客户细分框架

针对特定领域的企业采集其官方微博的粉丝及其微博好友的标签数据，利用文献[1]提出的客户特性表示方法构造客户特性描述文本；通过向量空间模型 VSM 和 TF-IDF 公式计算词语权重，将客户特性描述文本转换成文档–词语矩阵；利用不同文本聚类算法进行细分获得不同的簇；评价聚类结果，识别有意义的簇，并通过标签云可视化呈现结果；结合领域知识和专家分析，找出有助于微博营销的细分策略。

3.2　客户特性表示

微博用户可以自由定义标签以描述自我兴趣爱好，这些标签体现了用户在生活、职业等层面的特点，因而它们在一定程度上反映了用户的个人特性；同时，由于标签是用户自定义的，对用户个人的兴趣爱好等特性的描述更精炼、更准确。

用户的兴趣爱好受朋友同学等较为亲密的人影响；相反，朋友同学的兴趣爱好也在一定程度上反映该用户的兴趣爱好。类比到微博平台上，用户的社会关系在一定程度上体现在用户间的关注关系上，即用户会和其朋友同学形成互相关注的双向关系；用户也会主动关注感兴趣的媒体、领域名人和公共服务微博等，形成单向的关注关系。因此，融合用户及其微博好友的标签信息能从个人和社会关系两方面描述客户的特性，因此提出根据企业微博的粉丝(客户)及其关注的微博好友的标签生成客户特性描述文本的方法[1]。具体地，每个客户特性描述文本是由微博用户的标签及其关注好友的标签出现的总频数生成，其计算方法如下所示：

$$\text{userprofile}_i = \sum_{j=1}^{n} \text{friend}_j + \text{user}_i \qquad (1)$$

其中，user_i 表示用户 i 的标签向量，friend_j 表示用户 i 的好友 j 的标签向量(用户 i 总共有 n 个好友)，userprofile_i 则表示所得到的用户 i 的客户特性描述文本；用户的标签向量指的是以标签为维度、标签出现频数为维度值组成的向量。进而，每个客户特性描述文本可看做一个文档向量，并根据 TF-IDF 方法进行词语权重计算。

由于客户特性描述文本表示成的文本向量是高维稀疏数据，因此需要对高维文本数据进行有效的降维。

4　实验分析

4.1　实验数据

实验数据来自于新浪微博平台，以三个不同领域的企业官方微博为例，采集了企业官方微博的粉丝及其微博好友的标签，基本信息如表 1 所示：

表 1　企业官方微博账号基本信息

行业	微博账号名称	粉丝数	粉丝好友数	编号
旅游	完美旅行网	4 308	1 006 333	A
医疗健康	父母会育儿网	4 022	894 985	B
教育出国	澳洲留学辅导中心	5 000	686 545	C

为防止机器注册的用户(“僵尸粉”)造成噪声影响，本文根据“僵尸粉”的特征，设置条件对其进行移除。

通常，“僵尸粉”会通过持续关注不同用户形成大量的单向关注，同时“僵尸粉”的粉丝数量极少；且正常用户倾向于使用个性化域名(如 http://weibo.com/username)，而“僵尸”一般没有设置其域名。基于以上分析，以用户的相互关注数不少于10以及有定义个性化的用户域名为条件对“僵尸粉”用户进行剔除，筛选出质量较高的用户数据。处理后的数据基本信息如表2所示：

表2　移除“僵尸粉”后的企业基本信息

编号	粉丝数	粉丝好友数
A	1 073	337 241
B	1 714	295 139
C	2 691	328 240

4.2　实验分析过程

(1) 文本预处理

在实施文本聚类之前，需要对文本数据进行分词、去停用词、词频和文档频率统计、文本向量化等预处理。在本文实验中，文本预处理主要包括三方面的内容：

①基于用户标签构造用户特性表示文本，根据3.2节阐述的方法进行构造。

②考虑到传统降维方法的适用性和复杂性，本文主要通过两方面的处理实现简单的降维：针对标签存在大小写、繁简体等特点进行转换，过滤停用词；根据文档频率标准剔除在文档集合中出现频率高于90%、低于10%的词语。

③通过上述的步骤可得到由用户特性表示文本组成的文档集合，并依据TF-IDF权重计算公式，进而将用户特性表示文本进行向量化表示，得到文档–词语矩阵作为文本聚类算法的输入。

(2) 文本聚类过程

文本聚类过程涉及聚类算法选择、评价指标和参数设置等内容。

①聚类方法选择

本文选择K-means和层次聚类算法及基于NMF的聚类算法。K-means预先设置聚类数目K，将数据划分为K个簇。层次聚类算法则通过将数据组织为若干组并形成相应的树结构进行聚类。本文采用K-means和基于内平方距离法Ward的凝聚聚类算法，并采用适合文本数据的余弦相似度。

基于NMF的聚类算法主要有三个步骤：构造待分解的目标矩阵(本文指文档–词语矩阵)；对目标矩阵进行非负矩阵分解，得到由基向量组成的矩阵W和权重系数矩阵H；从分解后的矩阵中提取有意义的语义簇。

②聚类评价和参数K的选择

聚类的评价方式通常有外部和内部评价。外部评价是针对有标注类别的数据而言，有准确率、召回率和F值等指标；内部评价则针对没有标注的数据，通过计算簇内样本到簇中心的误差平方和来衡量簇内的凝聚性、计算簇间的距离总和来衡量簇间的分离性，以评估聚类效果的优劣，主要指标有Calinski-Harabasz Index(简称ch)[20]和Average Silhouette Width (简称asw)[21]。其中，ch是通过簇间距离平方和与簇内误差距离平方和的比值评价聚类效果，指标越大说明簇间距离相对较大、簇内距离较小，表明聚类效果较好，反之较差；asw指标则通过计算数据点与其所在的簇其他数据点、其他簇里的数据点的相异程度来衡量簇内凝聚性和簇间分离性，指标范围为[–1,1]，asw值越接近1表明效果越好，反之越差。由于实验数据缺少类别标签，同时考虑到三种聚类方法共有的评价指标，因此本文评价聚类效果时采用内部指标asw。

对于K-means和层次聚类算法中聚类数目K的选择，本文采取对不同的K值分别进行聚类，根据聚类评价指标ch和asw，选取聚类效果较好的K值作为最终的聚类数目。

针对NMF，需指定分解成的语义簇个数K和矩阵初始化算法。对于K的选择，采用Brunet等提出的方法[22]，通过不同K值对文本矩阵进行多次分解，构造一致矩阵(Consensus Matrix)，利用可视化重编码后的一致矩阵和共表型相关系数(Cophenetic Correlation)曲线图找到合适的K值。对于矩阵初始化算法，采用随机初始化的方式，通过多次迭代运行，减少分解的不稳定性。

以企业A的数据集为例，通过上述方法选取合适的参数。结合实际的营销知识，通常客户细分成的聚类数目不超过10个，因此在[2,10]内选择K值。由于ch和asw指标的取值范围有所不同(分别对应[–1,1]和[0,+∞])，为了便于观察曲线对应的指标最大值，对ch和asw进行规范化处理(数据集中的各项数据减去数据集的均值再除以数据集的标准差)。K-means聚类算法和层次聚类算法对应的K值和评价指标曲线如图2和图3所示。可以看出，K-means和层次聚类算法倾向于将数据划分成两个簇。

在非负矩阵分解K值的选取上，针对每个K值，通过50次非负矩阵分解，累加每次得到的连接矩阵(Connectivity Matrix)计算一致矩阵，重编码后绘制矩阵热图，如图4所示，从热图结构观察K值；并通过一致矩阵计算共表型相关系数以绘制曲线图，共表型系数衡量NMF分解后簇的稳定性，系数越大表明分解得到的一致矩阵更好。

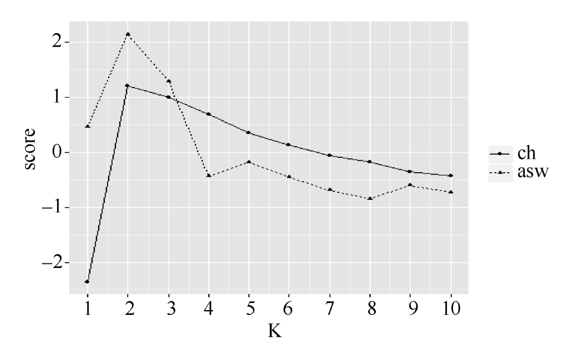


图 2　K-means 算法中 K 值选取

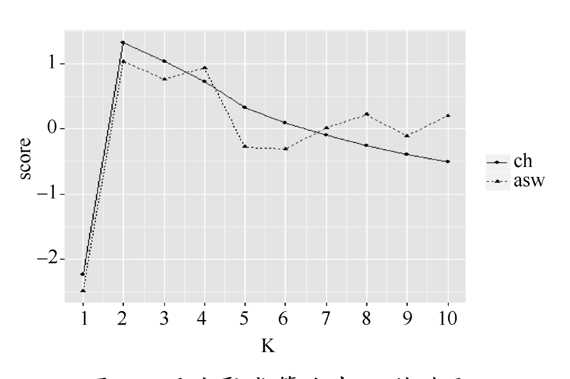


图 3　层次聚类算法中 K 值选取

从图 4 可看出 K 取 2 和 3 时，数据聚集成较大的深色方块，特别地当 K 取 3 时，数据形成 3 个深色的方块，表明从非负矩阵分解的角度看数据分为 3 个语义簇较为合理。当 K 取值为 4-10 时，可以看出数据开始趋向于聚集成多于 3 个的、不同程度的深色方块，但颜色分布不纯，说明数据中仍存在交叉重叠的语义簇，仍可继续细分；特别当K=8, 9, 10时，对角线上基本都有 8 个较为明显的方块，说明继续增大 K 值的矩阵分解倾向于形成 8 个方块，所以 K=8 可以作为另一个选择。结合共表型相关系数曲线图 5，可看出当 K=3 时，系数最大。综上，可以确定将文档–词语矩阵分解成 3 或 8 个语义簇。

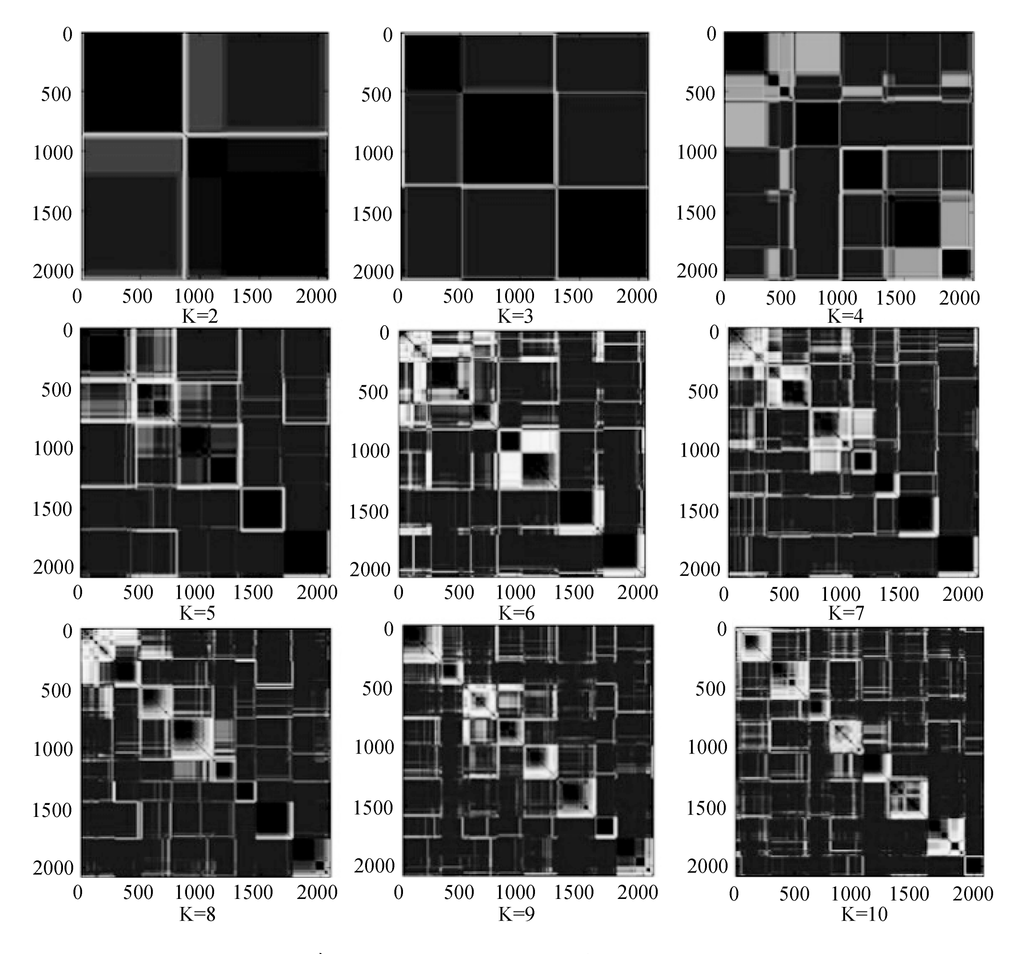


图 4　K 在[2,10]内取值时分别对应的一致矩阵热图

(注：在矩阵热图中，颜色值从 0 变化到 1。0 表示浅色，意味着数据样本不在同个簇内；1 表示深色，意味着数据样本分布在同个簇内。可以通过对角线的方块颜色和结构，观察 K 大致合适的数目。)

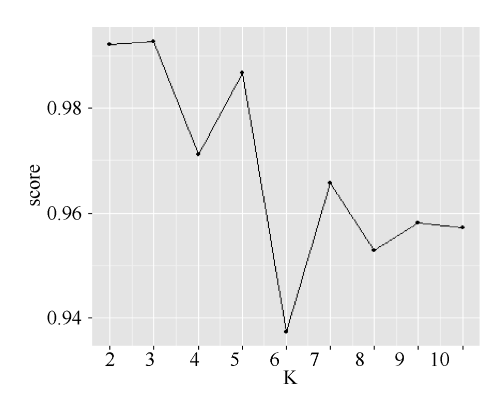


图 5　K 在[2,10]内取值时对应的共表型相关系数曲线图

同样地，针对企业 B、C 的数据集按照上述方法选择相应的 K 值，结果如表 3 所示：

表 3　不同企业微博数据对应的 K 值

企业编号	K-means/层次聚类选取的 K 值	NMF 选取的 K 值	
A	K=2	K=3	K=8
B	K=2	K=3	K=5
C	K=2	K=2	K=6

(3) 聚类结果分析和可视化

通过在不同行业的企业微博数据集上对比不同聚类算法的效果，利用三种聚类算法共同的评价指标 asw 进行评估；进而选择较好的聚类算法，对不同数据集进行聚类，通过标签云的形式可视化聚类结果。由于 K-means、层次聚类和 NMF 在 K 值选取上存在差异，因此本文针对不同算法选取的 K 值分别聚类并评价。

由表 4 看出，从评价指标方面，基于 NMF 的聚类算法远远优于 K-means 和层次聚类算法，以平均值来估算，基于 NMF 的聚类评价指标 asw 为 86.130%，远远超出基于 K-means 和层次聚类的方法。值得注意的是，当 K=2 或 3 时，基于 NMF 的聚类方法倾向于将文本数据粗略划分为 2-3 个簇，但从实际领域知识的角度看，这些划分仍比较粗糙，可进一步细分发现更有意义的簇，因此本文考虑 NMF 在 K 值选择方面的另一种方案如 K=8 或 5 或 6，虽然聚类评价指标可能有所降低，但有利于挖掘更有价值的客户群信息。

表 4　不同聚类算法在不同数据集、不同 K 值时的 asw 值

	A			B			C		平均值
	K=2	K=3	K=8	K=2	K=3	K=5	K=2	K=6	
K-means	0.03865	0.03827	0.04956	0.19138	0.13253	0.13988	0.04759	0.05290	0.08635
层次聚类	0.03402	0.03417	0.02457	0.19175	0.12785	0.12621	0.05018	0.03452	0.07791
NMF	1.00000	0.70000	0.70000	1.00000	0.78000	0.92000	1.00000	0.79000	0.86130

经过前面的分析，本文确定使用基于 NMF 的聚类算法对不同行业的文本数据进行聚类。以企业 A 为例，提取不同 K 值对应的细分结果，如表 5 和表 6 所示。

表 5　K=3 时对应的非负矩阵分解提取出来的簇关键词

簇标签	簇中按权重系数排序的前 15 个关键词
1 学生	学生、睡觉、上网、电子商务、交友、唱歌、动漫、篮球、创业、汽车、天蝎座、宅女、看书、运动、乐观
2 商务白领	电子商务、新闻、自由行、自驾游、户外、机票、度假、财经、自助游、传媒、创业、酒店预订、出境游、旅游达人、移动互联网
3 时尚爱好者	演员、美容、潮流、街拍、微时尚、搭配、女性、购物、爱情、创意、语录、淘宝、文艺、化妆、学生

从表5和表6可以看出，通过非负矩阵分解后实现的聚类，得到的语义簇还是比较有意义的。从表5可看出该企业微博客户群主要有学生、商务白领和时尚爱好者等三个群体，但这些簇仍可以继续细分，比如第二个簇可细分出旅游、汽车等客户群。由表6可以看到细分出来的客户群特征较表5更加具体、更有意义，对应了学生、时尚爱好者、商务白领、旅游爱好者、演艺人士、年轻妈妈、互联网从业者、创意艺术爱好者等。

类似地，对企业 B、C 对应的数据集进行聚类，通过标签云可视化客户群关键词，如图 6 和图 7 所示。可以看出企业 B 和 C 较为明显的客户群特征，分别对应的是母婴育儿和出国留学两种企业客户群特征。

表 6　K=8 时对应的非负矩阵分解提取出来的簇关键词

簇标签	簇中按权重系数排序的前 15 个关键词
1 学生	学生、睡觉、上网、唱歌、交友、篮球、动漫、宅女、天蝎座、运动、看书、乐观、狮子座、汽车、射手座
2 时尚爱好者	美容、潮流、女性、语录、爱情、学生、搭配、街拍、购物、化妆、星座、冷笑话、瘦身、减肥、淘宝
3 商务白领	新闻、财经、传媒、股票、杂志、金融、经济、汽车、历史、教育、房地产、投资、文化、主持人、商业
4 旅游爱好者	自由行、机票、度假、自驾游、出境游、酒店预订、自助游、户外、旅游达人、旅游业、签证、马尔代夫、欧洲旅游、旅游网站、国内游
5 演艺人士	演员、主持人、歌手、模特、艺人、音乐人、湖南卫视、明星、作家、天蝎座、香港、篮球、双鱼座、小说、导演
6 年轻妈妈	育儿、微时尚、母婴、亲子、早教、宝宝、淘宝、辣妈、妈妈、怀孕、购物、美容、代购、街拍、护肤
7 互联网从业者	电子商务、移动互联网、创业、营销、网络营销、产品经理、广告、汽车、手机、投资、财经、用户体验、风险投资、传媒、科技
8 创意艺术爱好者	创意、插画、漫画、摄影师、动漫、广告、杂志、视觉、设计师、美剧、文化、文艺、日本、动画、家居

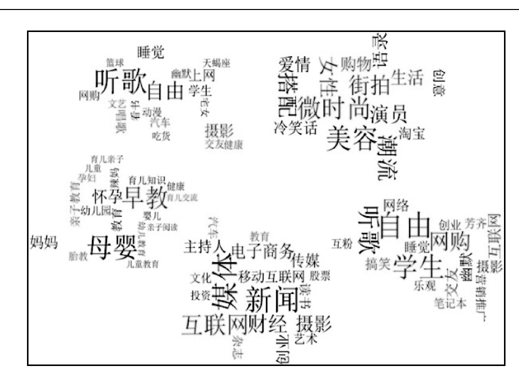


图 6　企业 B 对应的企业客户群关键词标签云

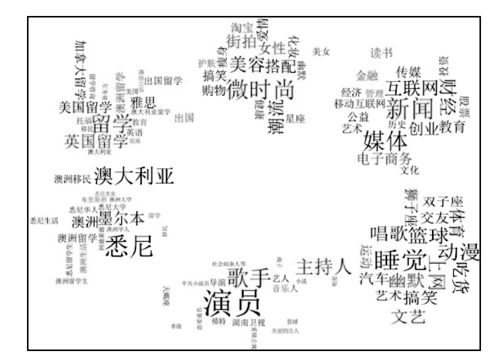


图 7　企业 C 对应的企业客户群关键词标签云

4.3　分析与讨论

从实验结果来看，基于 NMF 的文本聚类方法的评价指标大幅度超出了传统聚类方法对应的指标；同时从实际聚类效果来看，基于 NMF 的方法确实能够较好地发现不同的企业客户群特征。可以推测有以下两方面的原因：

(1) 传统文本聚类方法依据“文本中词与词之间互相独立”的假设，缺乏对语义的考虑；当面临高维稀疏的文本数据时，这些算法难以较好计算数据对象间的相似度以致无法有效实现聚类，故效果较差。

(2) 由于非负矩阵分解具有寻求降维表示和提取潜在语义的特性，因此将非负矩阵分解应用到文本聚类中，能挖掘文档集合的潜在语义；并通过分解后的基向量矩阵和权重系数矩阵的列向量加权组合来表示文档，直观解释文档语义；在 NMF 的基础上提取不同语义簇，间接实现文本聚类。通过多次迭代分解后可以得到精确的结果，所以基于 NMF 的聚类效果较好。另外，从图 6 和图 7 也可看出，细分的结果可能会出现极少数相似的语义簇。因此，从实际操作的角度，可考虑将客户细分为 5-8 个群体，再根据划分粒度的粗细进行适当的合并以得到更为合理的结果。

5　结　语

在社会化营销盛行的背景下，本文在客户特性表示的基础上，提出一种面向企业微博的客户细分框架。借助微博客户的个人和社会特性，利用微博客户及其微博好友的标签来表示该客户的特性；同时利用文本聚类技术对微博客户的标签文本进行聚类；在对聚类结果进行评价后并通过标签云的可视化呈现客户细分结果。实验结果表明，采用基于 NMF 的聚类算法明显优于传统的 K-means 和层次聚类算法，能使客户细分的结果更有意义。

然而，本文提出的框架相对简单，仍有以下方面需要完善：本文只通过融合微博客户个人及其关注好友的标签来表示微博客户的特性，尚未能全面刻画客户特征，后续可以结合微博客户注册的背景信息或微博文本，探索更好的微博客户特性表示方法；鉴于其他聚类算法的复杂性，本文仅考虑传统的 K-means 和层次聚类及基于 NMF 的聚类方法，可以继续探究其他算法的效果；只采用内部评价方式评估聚类，需要

结合领域知识以全面评估聚类效果；如何将提出的方法和框架应用到实际微博营销领域，也是未来的研究方向。

参考文献：

[1] Pang G S, Jiang S Y, Chen D Y. A Simple Integration of Social Relationship and Text Data for Identifying Potential Customers in Microblogging [A]. //Advanced Data Mining and Applications [M]. Springer Berlin Heidelberg, 2013: 397-409.

[2] Hennig-Thurau T, Malthouse E C, Friege C, et al. The Impact of New Media on Customer Relationships [J]. Journal of Service Research, 2010, 13(3): 311-330.

[3] Stelzner M A. Social Media Marketing Industry Report [EB/OL]. [2014-06-15]. http://www.socialmediaexaminer.com/SocialMediaMarketingReport2011.pdf.

[4] Rajagopal S. Customer Data Clustering Using Data Mining Technique [J]. International Journal of Database Management Systems, 2011, 3(4): 1-11.

[5] Lefait G, Kechadi T. Customer Segmentation Architecture Based on Clustering Techniques [C]. In: Proceedings of the 4th International Conference on Digital Society. IEEE, 2010: 243-248.

[6] Wu J, Lin Z. Research on Customer Segmentation Model by Clustering [C]. In: Proceedings of the 7th International Conference on Electronic Commerce. ACM, 2005: 316-318.

[7] Pennacchiotti M, Popescu A M. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter [C]. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 430-438.

[8] Tinati R, Carr L, Hall W, et al. Identifying Communicator Roles in Twitter[C]. In: Proceedings of the 21st International Conference Companion on World Wide Web. ACM, 2012: 1161-1168.

[9] Fink C, Kopecky J, Morawskib M. Inferring Gender from the Content of Tweets: A Region Specific Example [C]. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland. AAAI, 2012: 459-462.

[10] Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques [C]. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2000: 1-20.

[11] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review [J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[12] Willett P. Recent Trends in Hierarchic Document Clustering: A Critical Review [J]. Information Processing and Management, 1988, 24(5): 577-597.

[13] Rao D, Yarowsky D, Shreevats A, et al. Classifying Latent User Attributes in Twitter [C]. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. ACM, 2010: 37-44.

[14] Lee D D, Seung H S. Learning the Parts of Objects by Non-negative Matrix Factorization [J]. Nature, 1999, 401(6755): 788-791.

[15] Shahnaz F, Berry M W, Pauca V P, et al. Document Clustering Using Nonnegative Matrix Factorization [J]. Information Processing & Management, 2006, 42(2): 373-386.

[16] Wang X, Tang J, Liu H. Document Clustering via Matrix Representation [C]. In: Proceedings of the 11th International Conference on Data Mining. IEEE, 2011: 804-813.

[17] Gautam B P, Shrestha D. Document Clustering Through Non-Negative Matrix Factorization: A Case Study of Hadoop for Computational Time Reduction of Large Scale [J]. 稚内北星学園大学紀要, 2010, 10(3): 15-25.

[18] 黄钢石，陆建江，张亚非. 基于 NMF 的文本聚类方法[J]. 计算机工程，2004，30(11):113-114. (Huang Ggangshi, Lu Jianjiang, Zhang Yafei. Text Clustering Method Based on Non-negative Matrix Factorization [J]. Computer Engineering, 2004, 30(11): 113-114.)

[19] 张磊，冯晓森，项学智. 基于非负矩阵分解的中文文本主题分类[J]. 计算机工程，2009，35(13):26-27. (Zhang Lei, Feng Xiaosen, Xiang Xuezhi. Topic Classification of Chinese Document Based on NMF [J]. Computer Engineering, 2009, 35(13): 26-27.)

[20] Calinski T, Harabasz J. A Dendrite Method for Cluster Analysis [J]. Communications in Statistics, 1974, 3(1): 1-27.

[21] Rousseeuw P J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal of Computational and Applied Mathematics, 1987, 20(1): 53-65.

[22] Brunet J P, Tamayo P, Golub T, et al. Metagenes and Molecular Pattern Discovery Using Matrix Factorization [J]. Proceedings of the National Academy of Sciences (PNAS), 2004, 101(12): 4164-4169.

作者贡献声明:

陈东沂: 提出研究思路和实验方案;
周子程: 实验分析和验证;
周子程, 吴佳林: 采集、处理实验数据;
陈东沂, 周子程: 文献搜集和撰写论文;
蒋盛益, 王连喜: 论文修改和最终定稿。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, 可通过电子邮件向作者索取, E-mail: ziceweek@126.com。
[1] 周子程. data_V. rar. 微博大 V 关注与标签数据.

收稿日期: 2015-07-27
收修改稿日期: 2015-09-07

A Framework for Customer Segmentation on Enterprises' Microblog

Chen Dongyi[1,3]　Zhou Zicheng[1]　Jiang Shengyi[1]　Wang Lianxi[2]　Wu Jialin[1]
[1](School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006, China)
[2](Guangdong University of Foreign Studies Library, Guangzhou 510420, China)
[3](S.F.EXPRESS Co. Ltd., Shenzhen 518000, China)

Abstract: [**Objective**] This study tried to describe the customers' characteristics effectively. [**Methods**] The proposed framework aimed to explore the personal and social relationship among the customers and their friends on the microblog platform. We described the customers' characteristics using self-defined tags and then created segmentation with the help of text clustering and non-negative matrix factorization technologies. [**Results**] The method based on non-negative matrix factorization achieved an approximately 86.130% on average asw index, which outperformed traditional methods based on K-means and hierarchical clustering. [**Limitations**] The customers' characteristic cannot be described only by himself and his friends with self-defined tags on Microblogging. [**Conclusions**] The proposed framework could improve the effectiveness of characteristics description, evaluation and visualization of microblog customer segmentation.

Keywords: Customer segmentation　Microblogging marketing　Text clustering　Non-negative matrix factorization